Improving Recall in Appearance-Based Visual SLAM using Visual Expectation

Michael Milford and Gordon Wyeth
School of Engineering Systems
Queensland University of Technology
{michael.milford, gordon.wyeth} @qut.edu.au

Abstract

In this paper, we present a new algorithm for boosting visual template recall performance through a process of visual *expectation*. Visual expectation dynamically modifies the recognition thresholds of learnt visual templates based on recently matched templates, improving the recall of sequences of familiar places while keeping precision high, without any feedback from a mapping backend. We demonstrate the performance benefits of visual expectation using two 17 kilometer datasets gathered in an outdoor environment at two times separated by three weeks. The visual expectation algorithm provides up to a 100% improvement in recall. We also combine the visual expectation algorithm with the RatSLAM SLAM system and show how the algorithm enables successful mapping.

1 Introduction

Recently there has been an influx of state of the art visual SLAM systems that can create accurate maps of large environments in an online manner. The attraction of onboard visual sensors is of course their applicability in environments where GPS does not work, such as indoor or cluttered outdoor environments, as well as not needing to modify the environment by adding beacons or external camera systems. Furthermore, cameras have several attractive properties when compared with range sensors such as lasers; they can be very light, cheap, use minimal power and are passive sensors. The majority of visual SLAM systems developed thus far have been based around high quality stereo [1] or panoramic visual sensor data [2], although some researchers have focused on single camera systems [3, 4].[1]

In this paper, we focus on the other end of the visual sensor spectrum, by attempting to perform SLAM using 1000 pixel images and approximate self-motion information. Our appearance-based approach is intended for application in environments where robots tend to follow somewhat repeatable paths, such as indoor behaviour-based robots [5, 6] or perhaps in future applications on road-based vehicles [7]. We present a new visual expectation algorithm that enables recall to increase significantly without sacrificing precision, especially as the sensory data becomes more challenging. Using two 17 kilometre long datasets gathered three weeks apart on a suburban road network, we show that when combined with a lightweight visual odometry and visual template system, the expectation algorithm enables precise recall of places. We then combine the output of the visual expectation algorithm with the RatSLAM system to demonstrate successful mapping of the two datasets.

The paper proceeds as follows. In Section 2 we briefly review recent approaches to performing visual SLAM using stand alone cameras or robots. Section 3 provides a short overview of the RatSLAM SLAM system, which we use as the mapping backend to create the maps presented in this paper. The visual odometry, template, and expectation algorithms are described in Section 4. Section 5 provides details on the experimental setup, including the environment and dataset acquisition process. In Section 6 we present performance results identifying the effect of visual expectation on a single dataset and combined dataset consisting of two datasets obtained three weeks apart. Finally the paper concludes in Section 7 with discussion of the results and the identification of areas for future work.

2 Visual SLAM

The visual SLAM field is a highly active one in both robotics and computer vision, and there are a number of significant streams of investigation currently being pursued. Here we briefly mention some of the more seminal recent results in the field.

The FrameSLAM system uses bundle adjustment techniques to match visual frames based on point features in each frame, and stores information on the relative poses of frames [1]. It has been demonstrated mapping a 10 kilometer outdoor dataset in real-time. On a smaller scale, Davison [3] has achieved robust real-time SLAM using an Extended Kalman Filter based on real-time structure from motion. The MonoSLAM system with extensions has been applied in indoor and outdoor environments of up to 250 metres in length [8]. In work at Oxford on the FAB-MAP system, reliable online appearance-based mapping was

[1] This research was supported by an Australian Research Council Special Research Initiative on Thinking Systems, Grant ID TS0669699, to GW.

achieved over a 1000 km car journey on roads using high resolution panoramic images. Most recently, a 142 kilometre journey through Southern England was mapped into a relative map in an online manner using stereo data and bundle adjustment [9].

Lastly, we describe our research in brain-based and probabilistic visual SLAM. In past work we have presented the RatSLAM visual SLAM system, which is based on the neural processes underlying navigation in the rodent brain. RatSLAM has been demonstrated in a number of visual SLAM experiments [5, 7, 10]. The first was the mapping of a 66 kilometer journey through a suburban road network using only a single web camera as sensory input. This system used image intensity profiles as the primary input to the localization system, similar to those used by CMU's RALPH visual steering system [11]. The second was a 40 km long indoor delivery robot experiment, in which a robot performed SLAM and navigated to goal locations simultaneously at all times of the day over a period of two weeks. In this experiment, low resolution panoramic images were matched using a sum of absolute differences pixel matcher. Although RatSLAM is not the focus of this paper, we use it as a means of generating spatial maps and hence provide a brief overview in the following section.

Most recently, we have combined FAB-MAP and RatSLAM to demonstrate the potential for SLAM performance over multiple times of day [12]. In that work, we used images of sufficient resolution (640 × 480 pixels) for the FAB-MAP algorithm and underlying SURF features to be a tractable approach. In this paper, we use the same datasets but focus on the problem of performing scene recognition with low resolution images (1000 pixels), which is outside the normal operating range of classical feature-based techniques such as SIFT and SURF.

3 RatSLAM

The RatSLAM system consists of two major components – a continuous attractor neural network know as the *pose cells*, and a graphical map known as the *experience map*. The system requires two streams of sensory input – one which provides self-motion information, and one which provides some form of place recognition. In this section we briefly describe each of the RatSLAM components as implemented in this paper. More detailed descriptions of the RatSLAM system can be found in [7, 13].

3.1 Continuous Attractor Network

At the core of the RatSLAM system is a continuous attractor neural network of cells known as the *pose cell* network (Fig. 1). The pose cell network is structured as a 3D lattice grid of cell units, as shown in Fig. 1. Proximal cells are connected by both an excitatory and inhibitory 3D Gaussian distribution of weighted connections, which wrap across all three boundaries of the network. An iteration of the network's internal dynamics consists of local cell excitation and inhibition, followed by global inhibition and normalization of unit levels over the entire network.

3.2 Path Integration

Path integration is achieved by displacing the current activity state of the network by an amount proportional to the robot's translational displacement, in a direction corresponding to the angular orientation encoded by each cell. Because cells encode different robot orientations, a forward movement of the robot will result in cell unit activity propagating in different directions in the *(x′, y′)* plane of the pose cell network.

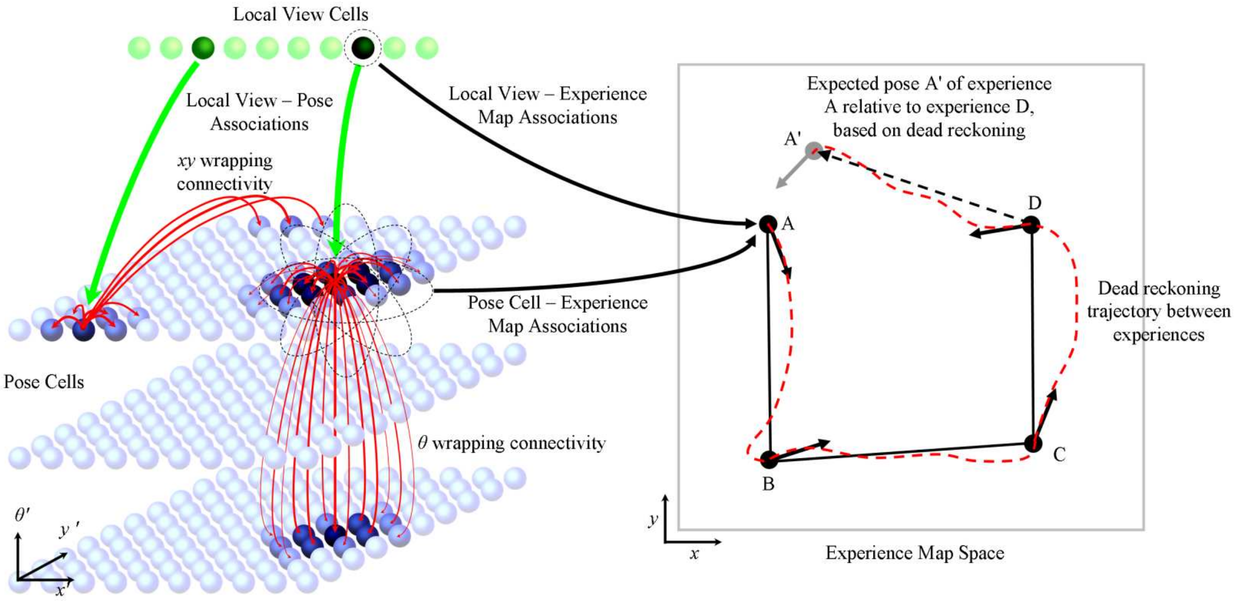
Local View Cells
Local View – Pose Associations
Local View – Experience Map Associations
xy wrapping connectivity
Pose Cell – Experience Map Associations
Pose Cells
θ wrapping connectivity
θ′
y′
x′
Expected pose A′ of experience A relative to experience D, based on dead reckoning
A′
A
D
Dead reckoning trajectory between experiences
B
C
y
x
Experience Map Space

Figure 1 – The RatSLAM pose cell network and experience map. Self-motion and visual data drive activity in the pose cell network, which in turn drives the creation of nodes and links in the experience map.

3.3 Local View Cells

The local view cells are an array of rate-coded units that pass information between the visual processing system and the pose cell network and experience map. As a robot explores an environment, each local view cell is allocated to a distinct visual scene in the environment. When the robot again sees that visual scene, or one similar to it, the local view cell is activated. In the simplified RatSLAM implementation described in this paper, only one local view cell can be active at any one time, and all cells' activity levels are binary on-off.

3.4 Experience Map

The experience map is a graphical map containing experience nodes, representing distinct places in the environment, and links between experiences, representing the movement transition between places (Fig. 1). Each experience node is associated with the state encoded in the pose cells and local view cells when it was created. New experience nodes are created when the pose cell and local view cell state does not match, within a threshold, the state encoded by any existing experiences. The experience nodes are themselves arranged in a Cartesian plane, and the map is continually adjusted through a process of graph relaxation.

4 Visual Algorithms

In this section we describe the visual odometry, visual template and visual expectation algorithms.

4.1 Visual Odometry

A lightweight visual odometry system was implemented using patch tracking of two fixed patch locations, shown in Fig. 2a. Each image frame was first resolution reduced to 320 × 240 pixels. The vehicle was treated as non-holonomic vehicle, with patch A used to track vehicle rotation, and patch B used to track the vehicle's translational speed. The comparison between patches was performed by calculating the average intensity difference, *f( )*, between pixel patches (normalized to 50% mean intensity) in the current and past image over a range of relative offsets:

$$f\left(\Delta x, \Delta y, I^j, I^k\right) = \frac{1}{r^2} \sum_{x=0}^{r} \sum_{y=0}^{r} \left(p^j_{x+\Delta x,\, y+\Delta y} - p^k_{xy}\right) \qquad (1)$$

where I^j and I^k are the past and current images, r is the patch size in pixels, p is the pixel intensity, and Δx and Δy are the patch offsets. The patch shift used for odometry purposes was the shift (Δx_m, Δy_m) that minimized *f( )* for the two patches:

$$\left(\Delta x_m, \Delta y_m\right) = \underset{\Delta x, \Delta y \in [-\rho, \rho]}{\arg\min} \; f\left(\Delta x, \Delta y, I^j, I^k\right) \qquad (2)$$

where ρ is the range of patch offsets. The horizontal pixel shift for patch A was multiplied by a gain constant, ς, to obtain a yaw velocity estimate, ω:

$$\omega = \varsigma \Delta x^A_m \qquad (3)$$

The gain constant ς was calculated using the camera's horizontal field of view. The net pixel shift for patch B was multiplied by a gain constant, v, to obtain a translational speed estimate, s:

$$s = -v\sqrt{\left(\Delta y^B_m - \Delta y^A_m\right)^2 + \left(\Delta x^B_m - \Delta x^A_m\right)^2} \qquad (4)$$

The translational speed gain constant, v, was calibrated on a separate set of video data. An example of the vehicle trajectory calculated using this lightweight visual odometry system is shown in Fig. 2b, and can be compared to the ground truth trajectory in Fig. 5. While the odometry method presented is not a general solution due to the scale ambiguity of monocular vision, it is sufficient in this application due to the relatively constant height of the camera above the usually flat groundplane. Furthermore, the experience map is not a globally metric map and hence does not require consistent metric odometry information.

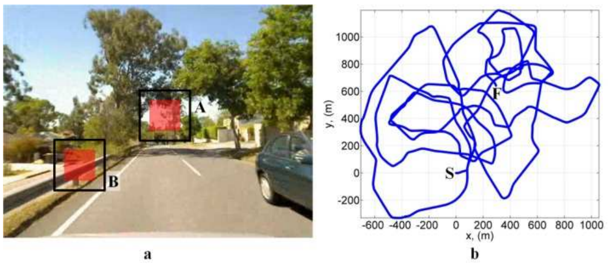

Figure 2 – Visual odometry illustration. (a) Patch locations and search ranges. (b) Trajectory for dataset 1 calculated using the visual odometry system.

4.2 Visual Templates

Each video frame was resolution reduced, converted to grayscale, and Gaussian blurred (radius 5) to form 32 × 32 images, which formed the basis of the visual templates used by the RatSLAM system (see Fig. 3). Template differences, D, between the current candidate template i and each template j were calculated using a normalized (normalized to a mean intensity of 50%) sum of pixel intensity differences performed over a moving sub frame in the resolution reduced images:

$$D_j = \min_{\Delta x, \Delta y \in [-\sigma, \sigma]} g\left(\Delta x, \Delta y, i, j\right) \qquad (5)$$

where σ is the template offset range, and *g( )* is given by:

$$g\left(\Delta x, \Delta y, i, j\right) = \frac{1}{s^2} \sum_{x=0}^{s} \sum_{y=0}^{s} \left(p^i_{x+\Delta x,\, y+\Delta y} - p^j_{x,y}\right) \qquad (6)$$

where s is the size of the template sub frame. These template differences were normalized by the current recognition threshold, T_j, of each template to calculate the template with the smallest normalized difference. The current template index, k, was calculated by:

$$k = \begin{cases} \underset{j \in [0,n]}{\arg\min}\, D_j / T_j & D_j / T_j < 1 \\ i & \min\left(\mathbf{D}/\mathbf{T}\right) \geq 1 \end{cases} \qquad (7)$$

where n was the number of learnt templates, and i was the index of the current template candidate. If no templates closely matched the current scene, the current candidate template *i* was added to the learnt templates. This same image difference metric was used to compare the current and immediately previous frame, to disable template learning and visual odometry for noisy corrupted frames.

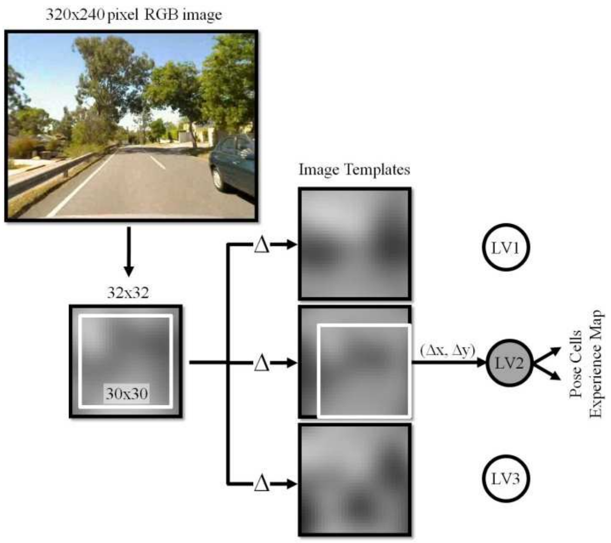

Figure 3 – Visual templates illustration. The resolution reduced current frame is compared against a library of image templates. If a template is matched, it activates the corresponding local view cell, otherwise a new visual template is learned.

4.3 Visual Expectation

In the mammalian brain, it has been shown that context can inform the recognition of objects or places. For example, the recognition of a car in a picture can be facilitated if a person knows that the candidate car object is located on a road – the road context increases the likelihood that the object will be recognized as a car [14]. Contextual recognition is achieved in the brain by dynamic adjustment of what are in effect recognition thresholds [15]. Based on this insight, we have developed a new visual expectation algorithm that dynamically modifies on an individual basis the recognition thresholds of visual templates. For example, if visual template A has in the past been followed immediately by template B, when template A is again recognized, the recognition threshold of template B is raised to increase the likelihood of it being matched. While contextual recognition has been explored extensively in the domain of object and scene recognition in computer vision [16], as well as by humans [14], its usage in recognition of scene *sequences* remains largely unexplored.

The visual expectation algorithm is implemented in the following way: the template comparison threshold, T_i, is adjusted as follows:

$$T_i = T_i + \sum_{j=i-\mu}^{i-1} \psi V_j - \alpha\left(T_i - T_D\right) \qquad (8)$$

where μ is the expectation range, ψ is the expectation increment per video frame, V is a binary array encoding the current template matches, and α is a per-frame threshold decay. T is bounded between a default threshold value T_D and a maximum threshold value T_m (set to twice T_D in these experiments). Figure 4 shows an example of how the recognition of a visual template can increase the recognition thresholds of subsequent templates and set off a sequence of recognized visual templates.

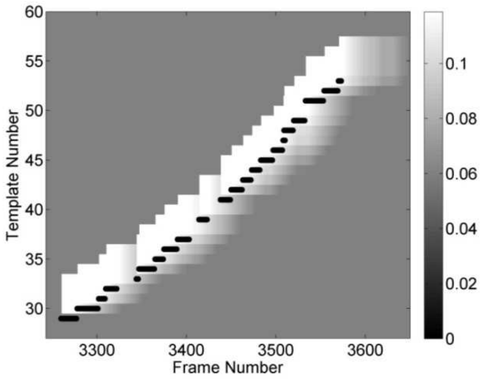

Figure 4 – Enhanced template recognition thresholds due to visual expectation. Dots indicate the currently matched template. The shading level indicates the threshold for visual template comparisons, below which two visual templates are considered to be the same. The lighter shaded areas above each matched template indicate the elevated recognition thresholds for templates immediately following a matched template. The elevated thresholds then decay over subsequent frames.

5 Experimental Setup

In this section we describe the testing environment, platform and data acquisition.

5.1 Platform and Data Acquisition

We used two datasets gathered using a Logitech QuickCam Pro 9000 web camera at 640 × 480 pixel resolution and an average frame rate of 15 frames per second. The camera was mounted on the top of a car windshield, facing forwards with a neutral pitch. Its field of view is approximately 62 degrees horizontally by 46 degrees vertically. To provide a ground truth measure, GPS positions were also logged at a frequency of 1 Hz. Each dataset consisted of approximately 25 minutes of driving over a distance of about 16.9 km. Both datasets were gathered at about midday, but dataset 1 was obtained 3 weeks after dataset 2.

5.2 Environment

The testing environment was a part of a suburban road network in Brisbane, Australia, shown in Fig. 5. This environment contains a highly varied range of terrain and scenery, including heavily vegetated sections and built urban environments. Since experiments were run during a normal weekday, there were constant changes to the environment in the form of moving vehicles and changing lighting conditions. Over the three week period separating the two datasets, the environment also changed due to these same factors, as shown in Fig. 6.

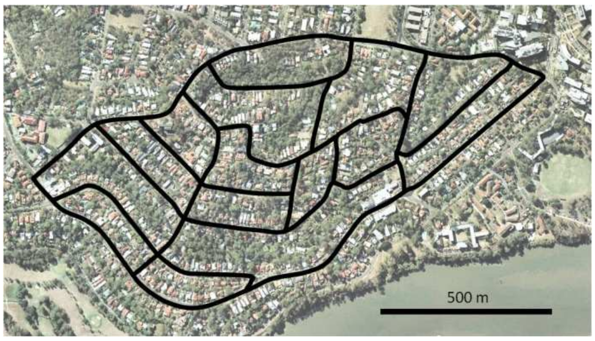

Figure 5 – Aerial photo of environment, including the path taken by the vehicle. (C) Google Maps.

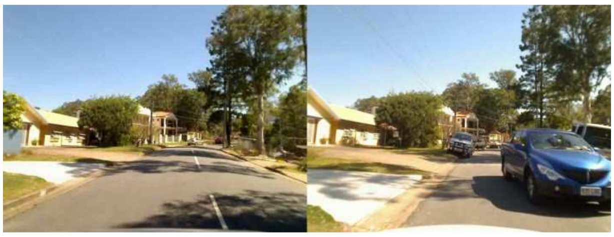

Figure 6 – Examples of frames from the two datasets.

5.3 Parameter Values

Table 1: Parameter Values

Parameter	Value	Description
r	32 pixels	Odometry patch size
ς	0.19 °/pixel	Yaw gain constant
ν	0.21 m/pixel	Translational speed constant
ρ	10 pixels	Patch offset range
μ	20 frames	Expectation range
s	30 pixels	Template sub frame size
σ	1 pixel	Template offset range
ψ	0.1	Expectation increment
T_d	0.01 – 0.15	Default threshold
α	0.02	Per-frame threshold decay
r_{thresh}	25 m	True positive distance threshold
θ_{thresh}	30°	False negative angle threshold

6 Results

In this section, we present the place recognition performance of the visual template system with and without visual expectation. The first comparison is performed using just dataset 1, while the second comparison uses both dataset 1 and dataset 2. We generate precision-recall curves, classification graphs overlaid on the vehicle's ground truth trajectory, and the learned and recalled visual template graphs. In addition, the experience maps for dataset 1 alone and dataset 1 and 2 combined are presented. Finally, we examine an illustrative visual template recall sequence with and without visual expectation.

Precision-recall graphs were generated by running 29 trials with and without visual expectation enabled (for a total of 58 trials) over a range of default template recognition thresholds (T_D). Each frame was classified as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). True positive frames were frames in which a visual template score, *s*, was below 1:

$$s = \frac{d}{r_{thresh}} + \frac{\theta}{\theta_{thresh}} \tag{9}$$

where *d* is the distance between the current frame's GPS location and the GPS location associated with the recalled template, and θ is the angular difference between the current frame's GPS orientation and the GPS orientation associated with the recalled template. r_{thresh} is a true positive distance error threshold, and θ_{thresh} is a true positive angular difference threshold. Using an orientation threshold ensured matches from significantly different directions of motion (such as at an intersection) were not expected. For scores of *s* larger than 1, the frame was classified as a false positive.

True negatives were frames where a new template was learned and there were no previously visited locations with an *s* score below 1. If there were previously visited locations with an *s* score below 1, the frame was classified as a false negative. All precision-recall plots have been truncated to exclude some points obtained from trials with extreme threshold values, corresponding to trials where only a few templates were learned for the entire environment. The 1 Hz GPS signal was interpolated to provide intermediate ground truth locations.

6.1 Single Day Dataset

Figure 7 shows the precision recall graph for dataset 1. With visual expectation, the peak precision-recall performance is at T_D = 0.045, with a precision of 98.5% and a recall rate of 91.6%. Without visual expectation, recall never reaches 91.6% (maximum recall is 90.5%), but a matching precision level is achieved at a recall rate of 85.0%. For this dataset, visual expectation enables a slight increase in recall performance at high precision levels (P = 80% to 99.5%). Figure 8 shows the frame classifications, with visual expectation enabled, superimposed on the GPS ground truth plot, for T_D = 0.045. For the false positive graph, lines connect the false positive frames with the location they erroneously recalled. All repeated sections of path were recalled, except for the occasional frame. There were a few false positive matches, but the error in each case was small, indicated by the lack of long error lines present. Figure 9 shows the frame classifications superimposed on the template graph. Most of the false negatives occurred at transitions between novel and familiar sections of path.

Figure 10 shows the experience map produced with visual expectation for dataset 1. The map contains 2777 experiences and 3082 inter-experience links. The squiggle at the bottom of the map is the start of the dataset. The map provides a coherent representation of the environment that is topologically correct and approximately metric at a local scale.

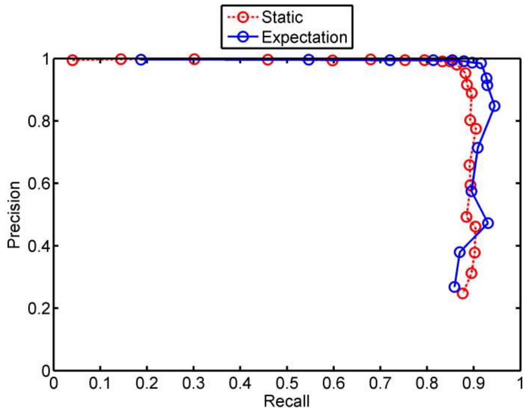

Figure 7 – Precision-recall graph with and without visual expectation for dataset 1. The lines between graph points indicate the direction of increasing default template recognition threshold, T_D.

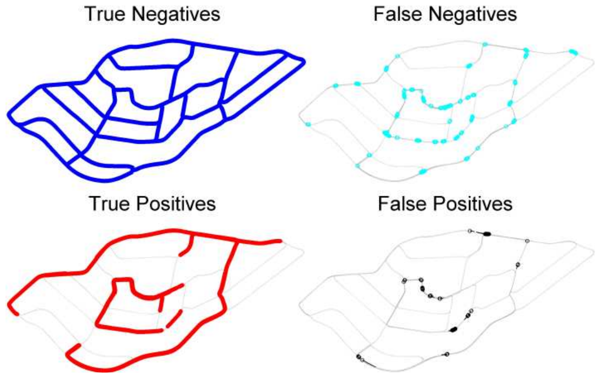

Figure 8 – Ground truth maps with true/false positive/negative classifications superimposed on the robot's trajectory for dataset 1, with visual expectation enabled. T_D = 0.045, P = 98.5%, R = 91.6%.

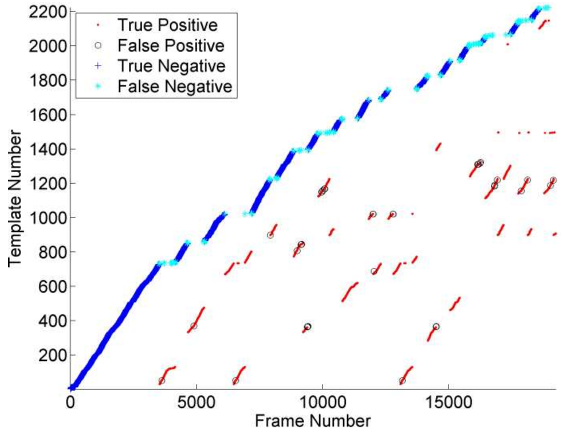

Figure 9 – Template graph for dataset 1 with visual expectation enabled, T_D = 0.045, P = 98.5%, R = 91.6%.

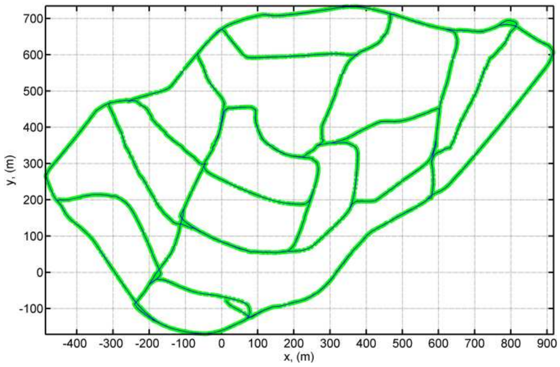

Figure 10 – The experience map created for dataset 1 with visual expectation (T_D = 0.045, P = 98.5%, R = 91.6%), containing 2777 experiences and 3082 links.

6.2 Three Week Interval Dataset

For the combined dataset consisting of dataset 1 and dataset 2, dataset 2 was treated as the "test" case – testing how well templates learned from dataset 1 were recalled in dataset 2. Consequently, there were no true negatives, as in theory every frame should have been recognized. All performance measures are given in terms of performance on dataset 2 after processing dataset 1.

For this combined dataset, visual expectation enables the precision to stay high at high recall rates, such as for T_D = 0.045, with a precision level of 97.0% and a recall rate of 78.7% (Fig. 11). Figure 12 shows the frame classifications, with visual expectation enabled, superimposed on the GPS ground truth plot for dataset 2, for T_D = 0.045. The majority of the path is recalled, although there are several sections where the system struggled to recall templates. It is also worth noting that visual expectation enables as high a recall rate as 89.8%, with a precision level of 89.7%.

Without visual expectation, the maximum recall rate achieved is 77.2%, however this is achieved at a markedly lower precision level of 72.5%. Figure 13 shows the consequences of the large number of false positives at this 72.5% precision level. To obtain a precision level of 97.0% or better without visual expectation, the recall rate drops to 40.0%. Fig. 14 shows the large number of false negatives resulting from such a low recall percentage.

Figure 15 shows the frame classifications with visual expectation enabled, superimposed on the template graph. While there are some false negatives, the majority of frames are successfully recognised. Figure 16 shows the experience map produced with visual expectation for both dataset 1 and 2 combined (T_D = 0.045). The map contains 3893 experiences and 4850 inter-experience links. Compared to the experience map for only dataset 1, there is a 40% increase in the number of experiences. This number is somewhat larger than the percentage of new visual templates seen in Fig. 15 because re-localization is not an instantaneous process – some new experiences are learned even as the system is re-localizing. It is also interesting to note the slightly squiggly nature of path sections on the left part of the map – when cross-referenced with the ground truth map shown in Fig. 12, it can be seen that this ‘squiggliness’ is most likely due to small recall errors between places along that path. There are a few sections of path where the repeated trajectories do not overlap, caused by the significant number of false negative matches, which can also be seen in Fig. 12. However the majority of the map still provides a representation of environment that would be usable for optimal path planning in robot navigation. The video accompanying this paper shows the evolution of the experience map during dataset 2, as well as the recall of visual templates from dataset 1.

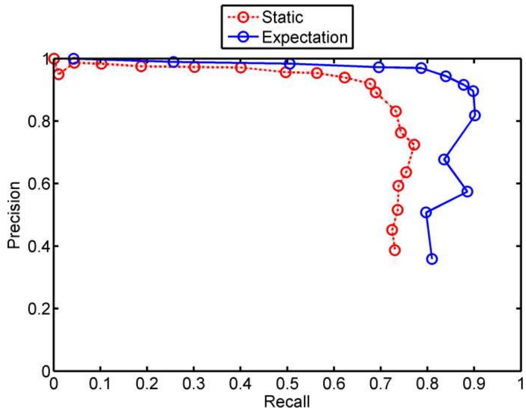

Figure 11 – Precision-recall graph for dataset 2 after processing of dataset 1, with and without visual expectation enabled.

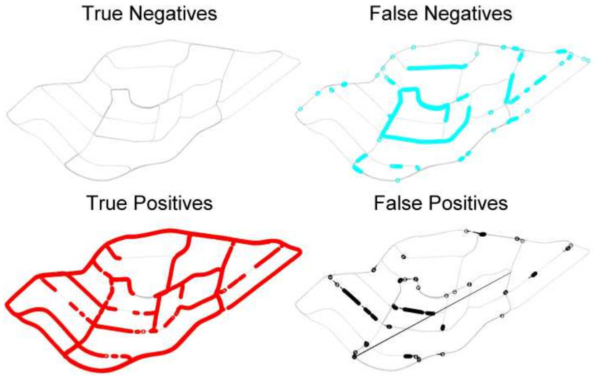

Figure 12 – Frame classifications superimposed on the GPS trajectory with visual expectation, T_D = 0.045, P = 97.0%, R = 78.7%.

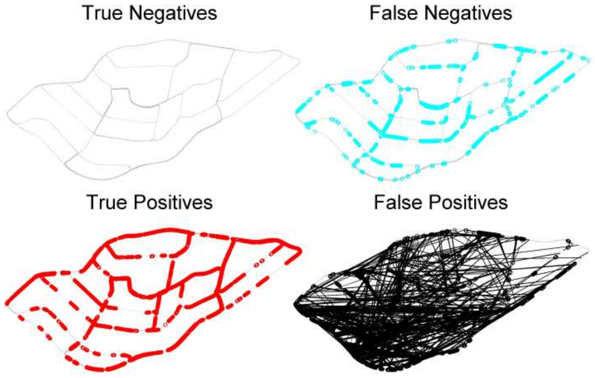

Figure 13 – Frame classifications superimposed on the GPS trajectory without visual expectation, T_D = 0.095, P = 72.5%, R = 77.2%.

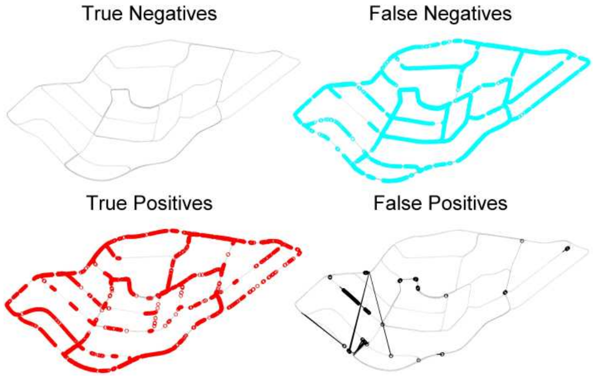

Figure 14 – Frame classifications superimposed on the GPS trajectory without visual expectation, T_D = 0.055, R = 40.0%, P = 97.1%.

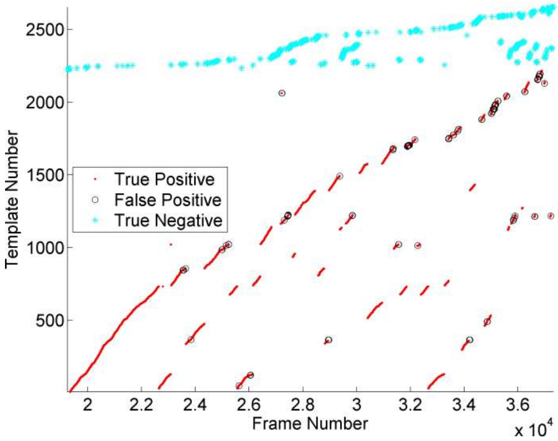

Figure 15 – Template graph for dataset 2 with visual expectation, T_D = 0.045, P = 97.0%, R = 78.7%.

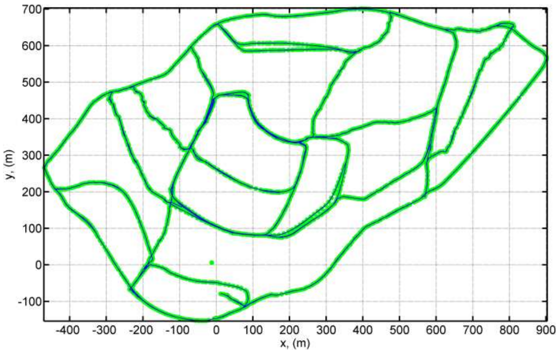

Figure 16 – Experience map for combined dataset 1 and dataset 2, containing 3893 experiences and 4850 inter-experience links, T_D = 0.045, P = 97.0%, R = 78.7%.

To visually illustrate the effect that visual expectation has, we lastly present examples of visual frame sequences through two sections of the environment at three week intervals. Next to the frame sequences, we show the visual template recall performance with and without visual expectation. Figure 17 shows a sequence of 13 frames at 25 frame intervals, and the recalled templates with visual expectation for T_D = 0.045, P = 97.0%, R = 78.7%, and without visual expectation for a matching a) precision level and b) recall rate. With visual expectation, a coherent and correct sequence of templates is recognized. Without visual expectation, to achieve the same level of precision, the system is only able to recognized a subset of the frames, with many false negative matches. When matching the recall level without visual expectation, there are several false positive incorrectly recognized visual templates, indicated by the black cross superimposed on the frame.

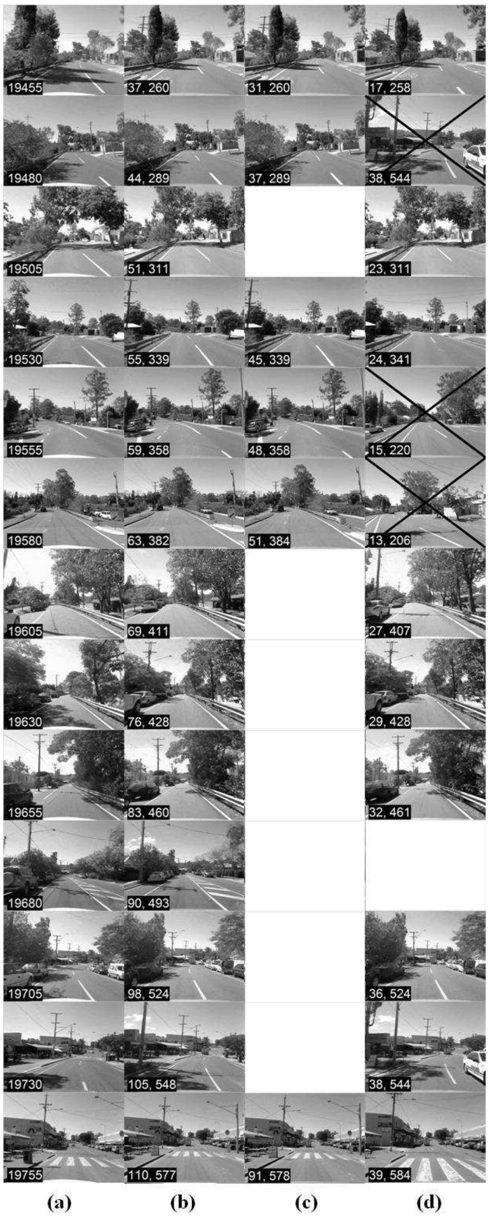

Figure 17 – (a) A sequence of frames from the second dataset, numbers indicate the frame number (b) Recalled visual templates with visual expectation, T_D = 0.045, P = 97.0%, R = 78.7%. Numbers are in the format (template number, associated frame number). Blank frames indicate false negatives, crossed frames indicate false positives. (c) Recalled templates without visual expectation for a matching precision level. (d) Recalled templates without visual expectation for a matching recall level.

7 Conclusions and Future Work

In this paper we have presented a visual expectation algorithm that increases the recall rates achievable while still maintaining high precision. The method provides a slight advantage on “ideal” data but becomes particularly effective on more challenging data, where it provides up to a 100% improvement in recall at high precision levels, and enables recall rates that are not achievable at any precision level without its use.

Future work will pursue a number of avenues of investigation. The current expectation algorithm assumes simple linear sequences of templates, which is valid for path-like datasets but becomes progressively less valid as the path divergence increases (i.e. a 5 road intersection). We will develop and test a more rigorous algorithm that can handle such situations as readily as single roads. Another weakness of the current algorithm is that it requires an initial template match to start a chain of template matches. We are currently investigating methods for coupling visual expectation with a method for matching sequences of weakly matching templates rather than just a single strongly matching template.

The flexibility of this algorithm will also be investigated by testing it on a range of visual datasets, such as the high resolution panoramic datasets used in FAB-MAP. Flexibility across platforms is currently being investigated, using datasets gathered from a quad-rotor flying platform. Initial results indicate the visual expectation method is still suitable even in environments where the camera path is less constrained and paths are not repeated exactly.

Acknowledgement

We thank Will Maddern and Arren Glover for their assistance in providing the datasets used in this paper.

References

[1] K. Konolige and M. Agrawal, "FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping," *IEEE Transactions on Robotics*, vol. 24, pp. 1066-1077, 2008.

[2] M. Cummins and P. Newman, "Highly Scalable appearance-only SLAM - FAB-MAP 2.0," presented at Robotics Science and Systems, Seattle, 2009.

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1052-1067, 2007.

[4] H. Strasdat, J. M. Montiel, and A. J. Davison, "Scale Drift-Aware Large Scale Monocular SLAM," presented at Robotics Science and Systems, Zaragoza, Spain, 2010.

[5] M. Milford and G. Wyeth, "Persistent Navigation and Mapping using a Biologically Inspired SLAM System," *International Journal of Robotics Research*, vol. 29, pp. 1131-1153, 2010.

[6] M. Milford and G. Wyeth, "Hybrid robot control and SLAM for persistent navigation and mapping," *Robotics and Autonomous Systems*, vol. 58, pp. 1096-1104, 2010.

[7] M. Milford and G. Wyeth, "Mapping a Suburb with a Single Camera using a Biologically Inspired SLAM System," *IEEE Transactions on Robotics*, vol. 24, pp. 1038-1053, 2008.

[8] L. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardos, "Mapping large loops with a single hand-held camera," presented at Robotics: Science and Systems, Atlanta, United States, 2007.

[9] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment," *International Journal of Robotics Research*, vol. 29, pp. 958-980, 2010.

[10] D. Prasser, M. Milford, and G. Wyeth, "Outdoor simultaneous localisation and mapping using RatSLAM," presented at International Conference on Field and Service Robotics, Port Douglas, Australia, 2005.

[11] D. Pomerleau, "Visibility Estimation from a Moving Vehicle Using the RALPH Vision System," presented at IEEE Conference on Intelligent Transport Systems, Boston, United States, 1997.

[12] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day," presented at International Conference on Robotics and Automation, Anchorage, United States, 2010.

[13] M. Milford and G. Wyeth, "Persistent Navigation and Mapping using a Biologically Inspired SLAM System," *The International Journal of Robotics Research*, 2009.

[14] A. Torralba, "Contextual influences on saliency," *Neurobiology of attention*, pp. 586–593, 2005.

[15] C. R. Nolan, G. Wyeth, M. Milford, and J. Wiles, "The race to learn: spike timing and STDP can coordinate learning and recall in CA3," *Hippocampus*, 2010.

[16] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," presented at International Conference on Computer Vision, Nice, France, 2003.